A Combined Stochastic and Deterministic Approach for Classification using Generalized Mixture Densities

Donald E. Waagen and John R. McDonnell

Abstract

This work investigates a combined stochastic and deterministic optimization approach for multivariate mixture density estimation. Mixture probability density models are selected and optimized by combining the optimization characteristics of a multiagent stochastic optimization algorithm based on evolutionary programming and the expectation-maximization algorithm. Unlike the traditional finite mixture model, generally composed of a sum of normal component densities, the generalized mixture model is composed of shape-adaptive components. Rissanen's minimum description length criterion provides the selection mechanism for evaluating mixture model fitness. The classification problem is approached by optimizing a mixture density estimate for each class. A comparison of each class's posterior probability (Bayes rule) provides the classification decision procedure. A classification problem is posed, and the classification performance of the derived generalized mixture models is compared with the performance of mixture models generated using normally distributed components. While both approaches produced excellent classification results, the generalized mixture approach produced more parsimonious density models from the training data.

1 INTRODUCTION

The area of nonparametric density estimation is proving to be an increasingly useful tool in providing a mathematical approach for the characterization and classification of complicated data. Several methods of nonparametric density estimation have been proposed, including (but not limited to) kernel estimators (Parzen 1962; Silverman 1986), maximum penalized likelihood estimators (Good and Gaskins 1971; Silverman 1982), and the method of mixtures (McLachlan 1986). Classification systems based on some neural network models, such as radial basis functions (Moody and Darken 1989) and the probabilistic neural network (Specht 1990), are mathematically and functionally equivalent to mixture models and kernel estimators, respectively. These neural network models therefore share the same characteristics and

19950719 051

limitations of their statistical analogues. The kernel estimator is simple to derive from the data, but requires the entire training sample for probabilistic inference. When compared to mixture models, kernel estimators have larger storage and longer run-time execution speed requirements. Finite mixture models provide data reduction and generalization, thus reducing the storage requirements and improving execution speed, but are computationally more expensive to derive.

This work investigates the use of combining stochastic search with the method of maximum likelihood for the optimization of mixture density estimates, where the number of components, the functional shape of each component, and the component parameters are simultaneously optimized. In this paper, these shape-adaptive mixture models are called *generalized mixtures*. The classification performance of generalized mixture densities is computed and compared to normal component mixtures for a two-class classification problem.

The following subsections introduce the mixture method approach for probability density estimation. The expectation-maximization (EM) algorithm for parameter optimization is described, and its relationship to finite mixture parameter optimization estimation is explained. Measures of model fitness and complexity are also discussed. Section 2 discusses stochastic approaches to optimization, including evolutionary programming. Section 3 discusses the formulation of the generalized mixture's components, and the combining of the EM and stochastic approaches for model order and parameter optimization. Section 4 describes a two-dimensional density estimation problem, and the performance characteristics of the stochastic-EM optimization process. Classification error rates of the evolved optimal generalized mixtures are computed and compared to normal-component mixture models in section 5. Conclusions are offered in section 6.

Finite mixture methods

A finite mixture distribution is defined informally as a distribution that decomposes into two or more proportionally scaled probability distributions. Mathematically, a mixture probability density function f, composed of q probability distributions $f_1,\ldots,f_q$, is defined as the following:

$$f(\mathbf{x}|\vec{\phi}) = \sum_{i=1}^{q} \alpha_i f_i(\mathbf{x}|\vec{\theta}) \qquad (1)$$

where $\vec{\phi}$ is the vector of free parameters $\vec{\phi} = [\vec{\alpha}, \vec{\theta}]^T$. The proportions $\alpha_1,\ldots,\alpha_q$ denote the relative contributions made by their respective density components. Their values are constrained by the following:

$$\sum_{i=1}^{q} \alpha_i = 1 \qquad \text{and} \qquad \alpha_i \geq 0 \quad (i = 1,\ldots,q). \qquad (2)$$

A goal of mixture model density estimation is to produce density estimates where the number of mixture components q is much smaller than the sample size n. Mixture models therefore attempt to provide some of the computational efficiency associated with parametric density estimation, while minimizing the number of assumptions concerning the true underlying distribution.

Unfortunately, analytical optimization of a finite mixture is complicated even for a moderate sample size (Choi and Bulgren 1968). A nonanalytical solution for mixture model optimization is provided by the EM algorithm.

Expectation-maximization algorithm for finite mixture optimization

An important generalization of the method of maximum likelihood was developed by Dempster et al. (1977), in which an iterative procedure was introduced that allows maximum likelihood estimates to be generated from incomplete data. Here "incomplete" is used in the sense that a component of the data is occasionally (or always) missing, and therefore some data does not provide values for the variables under consideration.

Suppose it is desired to find the maximum likelihood estimate of θ for the likelihood function $L(\theta) = g(x|\theta)$, where x is a set of "incomplete" data. Let y be a complete version of the data, and let the likelihood of y be denoted as $f(y|\theta)$. From an initial approximation $\theta^{(0)}$, the EM algorithm generates an iterative sequence of estimates $\theta^{(k)}$ via the following two steps:

$$E \text{ step: } \textit{Compute } Q(\theta|\theta^{(k)}) = E\left[\{\log f(y|\theta)\}|x, \theta^{(k)}\right] \tag{3}$$

$$M \text{ step: } \textit{Set } \theta^{(k+1)} = \max_{\theta \in \Omega} Q(\theta|\theta^{(k)}) \tag{4}$$

In solving a finite mixture problem, where the number of components is determined a priori, Redner and Walker (1984) give a formulation for the general EM approach. For a given vector of parameters $\phi^c = (\alpha_1^c, \ldots, \alpha_q^c, \theta_1^c, \ldots, \theta_q^c)^T$, which is the current approximate maximum likelihood estimate of the log-likelihood function $\log L(\phi|x)$, the next approximate maximum likelihood estimate $\phi^+ = (\alpha_1^+, \ldots, \alpha_q^+, \theta_1^+, \ldots, \theta_q^+)^T$ of the log-likelihood function is given by

$$\alpha_i^+ = \frac{1}{n} \sum_{k=1}^{n} \frac{\alpha_i^c p(x_k|\theta_i^c)}{p(x_k|\phi^c)} \tag{5}$$

$$\theta_i^+ \in \arg\max \sum_{k=1}^{n} \log p_i(x_k|\theta_i) \frac{\alpha_i^c p(x_k|\theta_i^c)}{p(x_k|\phi^c)}. \tag{6}$$

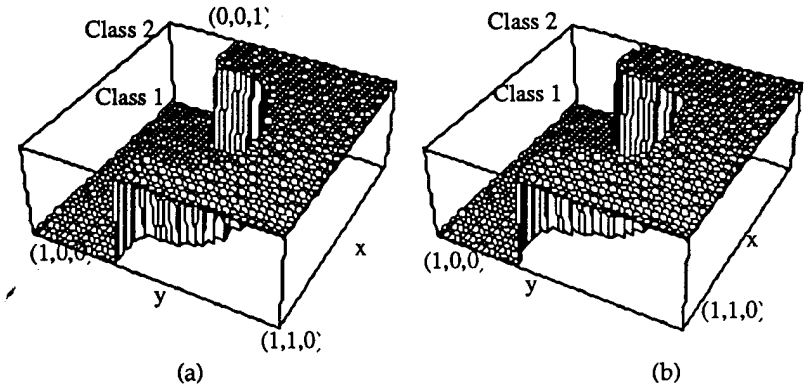


Figure 10. Generalized mixture model (a) and traditional normal mixture (b) decision surfaces for the two-class Flick data problem.

Note that the term $\alpha_i^c p(x_k|\theta_i^c)/p(x_k|\check{\phi}^c)$ is the posterior probability that x_k originated from the ith component population, given the current maximum likelihood estimate ϕ^c.

At each iteration, the EM algorithm guarantees that $\log L(\phi^+|x) \geq \log L(\phi^c|x)$. Given an a priori selection of the number of mixture components, this approach iteratively determines the proportional and parametric components of each component. An information criterion relating density estimate fitness and model complexity is required for comparison of mixture density estimates with a varying number of components.

Mixture model order identification

Models are generally compared by their complexity and how well they fit the data, with the goal of maximizing model fitness while minimizing model complexity. These constraints generally are diametrically opposed, i.e., increasing model complexity will generally allow for increasing the model fitness. To measure how well the models fit the data, a function incorporating the likelihood function is generally used. The likelihood function has several desirable properties which make it appropriate for measuring the relationship between model and data. Cramer (1946) demonstrates that under certain regularity conditions, the maximum likelihood estimators (MLE) of a multivariate model have solutions which are asymptotically normal and joint asymptotically efficient estimates of the parameters. Therefore, the maximum likelihood estimates of a set of parameters have the smallest variance about the true parameter values for all unbiased estimators. The MLE thus provides a very sensitive measure of fitness between model and data.

Many criteria of model complexity determination have been developed from the concept of maximizing the likelihood function while penalizing the number of free parameters required. These criteria include the generalized likelihood ratio test (Casella and Berger 1990), Akaike's information criterion (Akaike 1974), and the minimum description length (MDL) criterion (Rissanen 1986). The MDL criterion is formulated as

$$MDL(x) = \min_{k,\theta}\left\{-\log L(\theta|x) + \frac{k}{2}\log n\right\} \qquad (7)$$

where $L(\theta|x)$ is the likelihood function of the model, k is the number of the free parameters used to represent the data, and n is the cardinality of x. It is interesting to note that Schwarz's criterion (Schwarz 1978), a Bayesian derived criterion for model selection, is identical to the MDL criterion. Given a data sample x, a density estimate model M_i is "better" than another model M_j if $MDL(x|M_i) < MDL(x|M_j)$. In the

present work, the MDL criterion provides the likelihood-based measure of model fitness in the mixture model optimization process.

2 STOCHASTIC OPTIMIZATION

Random (or stochastic) search techniques have been used for function optimization since the 1950s. Stochastic search strategies are competitive with or superior to traditional search strategies (such as gradient search techniques) when the cost or objective function under optimization is difficult to compute, or when the function to be minimized has many suboptimal solutions (local minima). Other advantages, enumerated by Karnopp (1963), include the ease of programming, inexpensive realization of possible solutions, as well as flexibility in the expression of the criterion function.

Stochastic optimization techniques are based on either single point or multiple agent algorithms. Single point algorithms include the random walk, the creeping random method (Brooks 1958), and the method of Solis and Wets (1981). Multiple agent stochastic search algorithms, such as genetic algorithms (Goldberg 1989), evolution strategies (Bäck and Schwefel 1993), and evolutionary programming (Fogel 1991), are becoming well known for their optimization properties.

Figure 1 is a graphical representation of an evolutionary programming (EP) algorithm. A population of models (solutions) generates new models (offspring) via mutation, and then the complete population competes for survival to the next iteration. In EP, mutation consists of random perturbations of the parameters, with the magnitude of the perturbation generally self-adaptive or tied to the fitness of the parent. For the process of real-valued parameter optimization, the mutation process generally consists of perturbing a parameter value with a normal or lognormally distributed random variable. The normal perturbation scheme is given as the following:

$$\theta^{+} = \theta + N(0, f \cdot s + z) \tag{8}$$

where f is the measure of error of the parent model, and z and s are an offset and scale factor, respectively. The random nature of this mutation, although useful in escaping local minima, is inefficient for direct minimization. Also, the use of the error value f in the variance term $f \cdot s + z$ allows the variance to shrink as the error decreases (model fitness improves), but is properly defined only when $f \cdot s + z > 0$. Several approaches have been developed to avoid any association of the fitness value with the mutation process, and to speed the convergence to optimal solution (Waagen et al. 1992; McDonnell and Waagen 1994). The next section describes the hybrid approach developed in this work for mixture model selection and optimization.

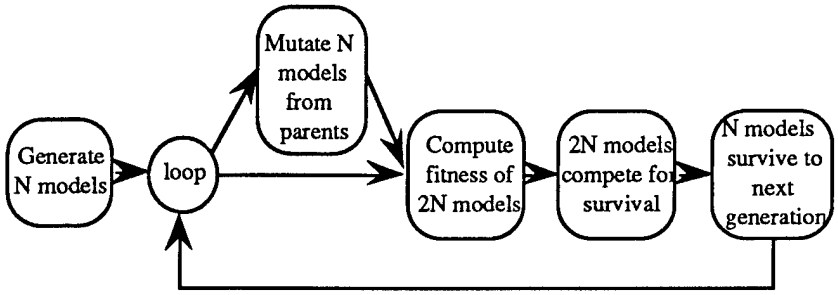


Figure 1. Evolutionary programming optimization algorithm. The mutation and competition steps are stochastic in nature.

3 STOCHASTIC-DETERMINISTIC MIXTURE DENSITY ESTIMATION

This section details the process of determining a multivariate mixture model from a set of independent, identically distributed data. Preceding the discussion of the optimization process, a discussion of the distributional form of the components which make up the mixture is in order.

Generalized mixture component formulation

The functional form of the mixture component distribution is based on a generalized kernel function described by Fukunaga (1990). The functional form of the generalized mixture components, denoted as $f(x|\mu,\Sigma,m)$, is given as

$$f(x|\mu,\Sigma,m)=\frac{m\Gamma\left(\frac{n}{2}\right)\Gamma^{n}\left(\frac{n+2}{2m}\right)}{(n\pi)^{n}\Gamma^{n+1}\left(\frac{n}{2m}\right)}\times\frac{1}{|\Sigma|^{n}}\times\exp\left[-\left\{\frac{\Gamma\left(\frac{n+2}{2m}\right)}{n\Gamma\left(\frac{n}{2m}\right)}(x-\mu)^{T}\Sigma^{-1}(x-\mu)\right\}^{m}\right] \tag{9}$$

where n is the dimensionality of the data, μ is the mean, Σ is the covariance matrix, and m is the shape parameter of the mixture component. The probability density function of the q-component mixture distribution is therefore given as

$$f(x)=\sum_{i=1}^{q}\alpha_i f(x|\mu_i,\Sigma_i,m_i) \tag{10}$$

The shape parameter m allows the function to include both the multivariate normal and multivariate uniform distributions as special cases. A graph of the univariate distribution of $f(x|0,1,m)$ for two values of m is given in Figure 2.

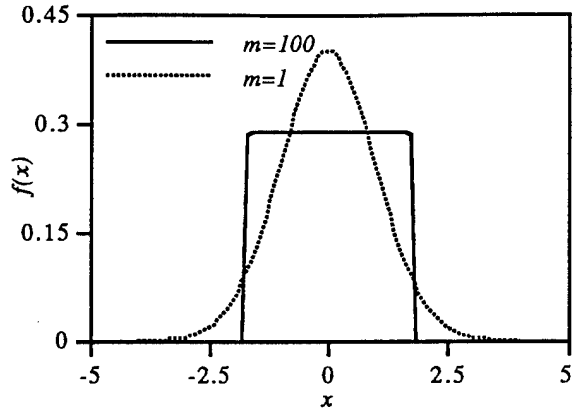


Figure 2. Distribution of $f(x|0,1,m)$ from Eq. 9 for m = 1, 100. Special cases of Eq. 9 include the multivariate uniform and normal distributions.

Hybrid multiagent approach to optimization

The multiagent optimization process investigated in this work combines a stochastic mutation process with the deterministic EM algorithm for component number and parameter optimization. A population of N possible solutions (mixture models) is generated and maintained throughout the optimization process. Optimization consists of two steps, as graphically shown in Figure 3. Optimization ends either after a fixed number of iterations or after the best mixture's error value remains constant for a significant number of iterations.

One issue with the combined stochastic-deterministic optimization approach is the fact that the parameters of a mutated offspring are not optimized for its component or shape representation, whereas the parameters values of the offspring's parent have been optimized (to some extent) for the mixture's shape and component number. Therefore the offspring models are initially at a disadvantage to the parent models in the competition process. An attempt to alleviate this problem is made by allowing each offspring model's parameters to be optimized via the EM algorithm for several iterations before model selection occurs. This optimization might be biologically analogous to the environmental learning phase of childhood and adolescence, but this paper will not try to justify or defend this analogy as truth.

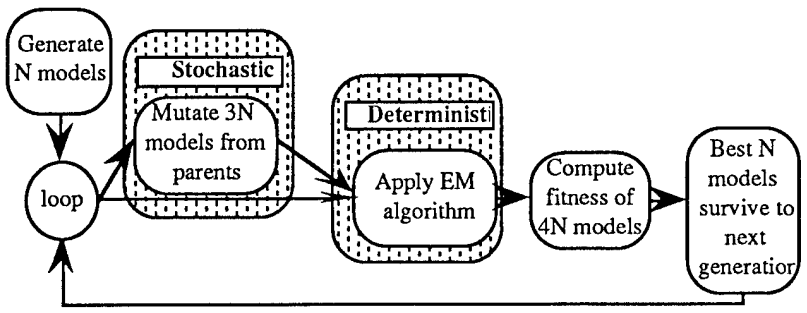


Figure 3. Multiagent mixture optimization procedure. The stochastic and deterministic portions of the approach are correspondingly labeled. Note that the competition is deterministic, and the best N models are always selected for survival.

The first step is mutation, consisting of the generation of three mixture models from each surviving parent model. The goal of mutation is to optimize of the number of components in the mixture and the shape parameters of each component. Three new models (offspring) are produced via the following rules:

Offspring 1: Add or remove a randomly selected component.

Offspring 2: Modify the shape parameter of a randomly selected component via the positive value of a random normal perturbation:

$$m^{+} = \begin{cases} m + |N(0,c)| & \text{if } m + |N(0,c)| \le M_{high} \\ M_{high} & \text{if } m + |N(0,c)| > M_{high} \end{cases} \tag{11}$$

where M_{high} is an arbitrary constant (200 in this investigation).

Offspring 3: Modify the shape parameter of the same selected component (as offspring 2) via the negative value of the same (as offspring 2) random normal perturbation:

$$m^{+} = \begin{cases} m - |N(0,c)| & \text{if } m - |N(0,c)| \ge 1 \\ 1 & \text{if } m - |N(0,c)| < 1 \end{cases} \tag{12}$$

After the offspring models have been generated and the EM algorithm has been applied repeatedly to each child, the combined population of models is individually passed through a single iteration of the EM algorithm. In each model, the EM algorithm optimizes the mean and covariance parameters, as well as the component proportion values, according to equations (5) and (6). The fitness of each model is then computed using the MDL (7), the best N models (from the population of $4N$ models) are kept, and the process is repeated.

4 MIXTURE MODEL OPTIMIZATION CHARACTERISTICS

To test the capability of the optimization process, a two-dimensional classification problem is posed. Two-dimensional problems aid in visualization and interpretation. Barring sample size issues, the generalized mixture technique is directly applicable to higher

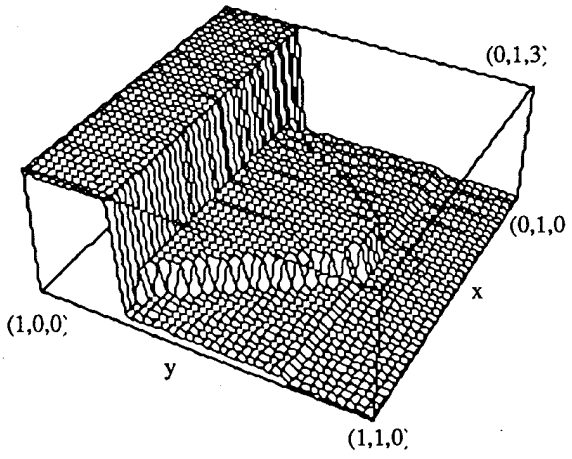
dimensional problems. In the experiment, training samples are created from each class's underlying distribution. These samples are used to derive generalized mixture models for each class. For comparison purposes, optimal (in the MDL sense) normal-component models are also derived for the data.

Flick data

To test the capability of the density estimation process, and to assess the classification capability of the generalized mixture model approach, two overlapping, piecewise-continuous distributions previously described by Flick et al. (1990), (herein labeled as classes 1 and 2) are estimated via the mixture distributions discussed in section 3. The probability density functions for class 1 [class 2] are defined on the unit square as the following:

$$f(x,y)=\begin{cases} 3.0 \quad [0.0] & 0 \le y \le 0.25 \\ 0.86 \quad [0.14] & 0.25 < y \le b(x) \\ 0.14 \quad [0.86] & b(x) < y \le 0.75 \\ 0.0 \quad [3.0] & 0.75 < y \le 1.0 \end{cases} \tag{13}$$

where $b(x)=0.5-0.25\cos(2\pi x)$. Figure 4 graphically displays the underlying probability density function for each class. These distributions overlap, so that the optimal classification rule, based on knowledge of the true underlying probability distribution of each class, will misclassify (on average) 3.5% of the new data sample presented for classification (3.5% of each class).



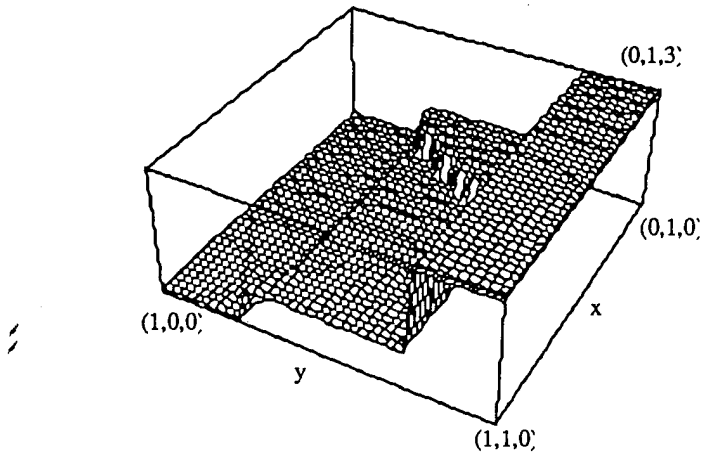


Figure 4. Probability density functions for classes 1 (top) and 2 (bottom).

The training sets consist of 200 samples of each class. The training sample for each class is shown in Figure 5. For the model optimization process, each class was optimized separately, using the MDL as the measure of model fitness. The model population number N (of Figure 3) was arbitrarily set to 5.

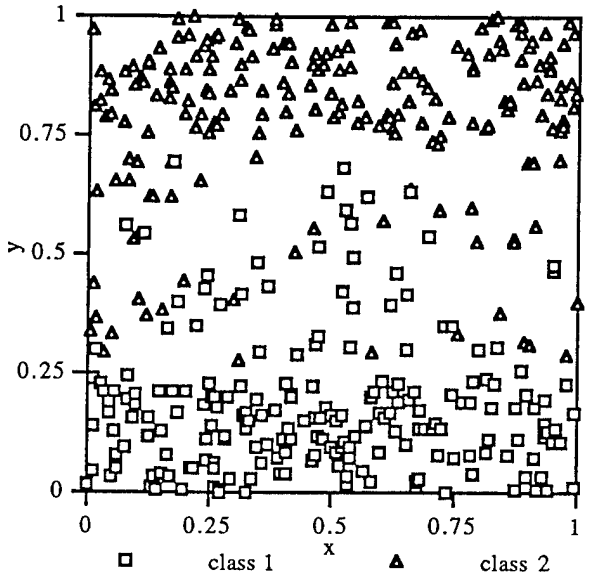


Figure 5. Training sets of the two-class problem. Each sample consists of 200 points.

For both classes, the algorithm converged to a solution within 50 iterations. The value of the MDL for the "best" mixture model and its corresponding number of components at each iteration are shown in Figure 6. Table 1 and Figure 7 display the resulting mixture models

derived by the algorithm from the sample data. The results demonstrate that the algorithm quickly minimizes the number of components in its optimization process, with both data samples being optimally modeled by two-component mixture models.

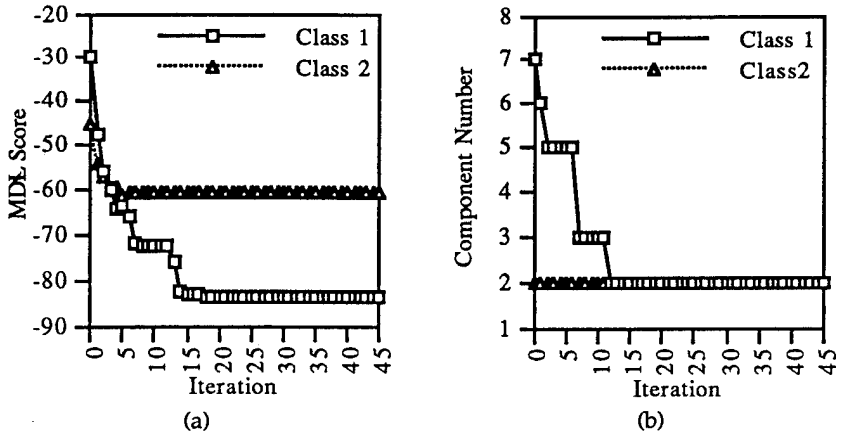


Figure 6. Best model's MDL score (a) and number of components (b). Model optimization occurs quickly, with both density models optimized within the first 20 iterations.

Table 1. Final generalized mixture model estimates for each class derived via stochastic-deterministic optimization. Each model was derived from 200 data points.

Class	Mixture Estimate $\sum_i a_i f(x\|\mu_i, \Sigma_i, m_i)$
1	$0.7805 \cdot f\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} 0.5146 \\ 0.1354 \end{bmatrix}, \begin{bmatrix} 0.08035 & -0.00179 \\ -0.00179 & 0.00554 \end{bmatrix}, 2.172\right) +$ $0.2195 \cdot f\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} 0.4790 \\ 0.4668 \end{bmatrix}, \begin{bmatrix} 0.04007 & -0.00362 \\ -0.00362 & 0.01960 \end{bmatrix}, 1.344\right)$
2	$0.6628 \cdot f\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} 0.5303 \\ 0.8780 \end{bmatrix}, \begin{bmatrix} 0.08399 & 8.3E-4 \\ 8.3E-4 & 0.00582 \end{bmatrix}, 4.424\right) +$ $0.3372 \cdot f\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} 0.5193 \\ 0.6394 \end{bmatrix}, \begin{bmatrix} 0.10910 & -0.00861 \\ -0.00861 & 0.03573 \end{bmatrix}, 4.070\right)$

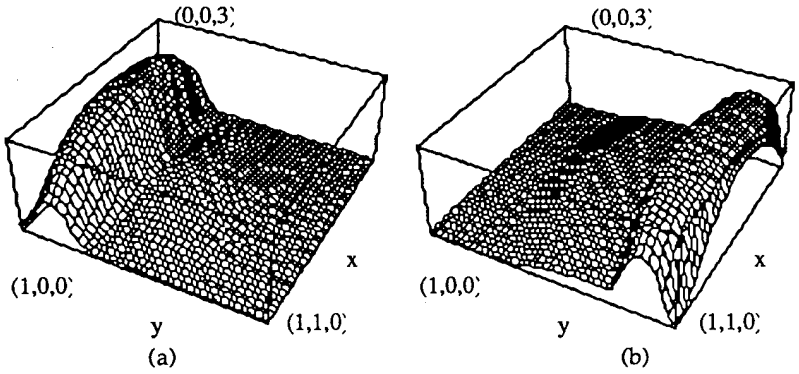


Figure 7. Generalized mixture models (class 1 (a), class 2 (b)) generated by stochastic-deterministic optimization.

Normal component mixture model comparison

To investigate its utility, the generalized mixture approach is compared with the traditional mixture approach. The traditional mixture consists of normally distributed components. Computing models using normal

components illuminates the representation and classification capability of the generalized mixture models.

Optimal (in terms of the mixture's MDL value for the training data set) normal component mixture density estimates were computed from the training data. Figure 8 display the normal mixture models of each class.

The densities generated via the two approaches are given in Table 2. As shown in the table, the generalized mixture approach produces superior models in terms of the log-likelihood function, the total number of free parameters used for data representation, and the MDL criteria. This superiority is due to the added flexibility of component shape modification. The next section compares the classification capability of these density models.

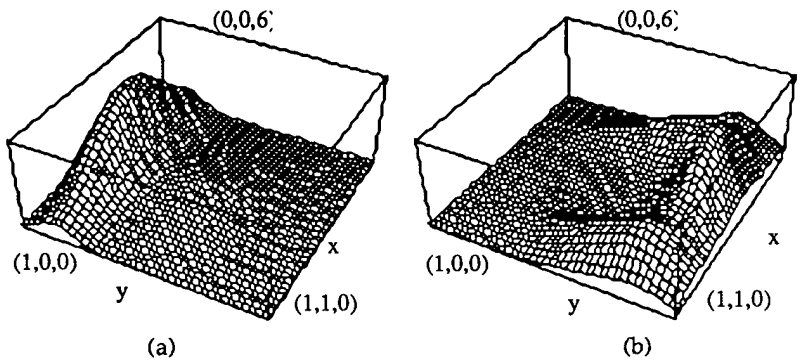


Figure 8. Optimal normal component mixture models (class 1 (a), class 2 (b)).

Table 2. Comparison of fitness characteristics of generalized and normal mixture models derived from the training data. The models are optimal with respect to the training data and the criterion function, the MDL (Eq. 7).

Feature	Generalized Mixture		Normal Mixture	
	Class 1	Class 2	Class 1	Class 2
Number of components	2	2	2	3
Log-likelihood function	115.585	95.274	106.177	94.270
Number of free parameters	13	13	11	17
MDL score	-81.5104	-60.8346	-77.0359	-49.2344

5 CLASSIFICATION RESULTS

The mixture models produced are probability density functions, so it is natural to use Bayes rule as the decision rule for classification of new

data. Given a set of c possible classes ω_i, Bayes rule is given as the following:

$$P(\omega_i|x,\theta_i)=\frac{f(x|\omega_i,\theta_i)p(\omega_i)}{\sum_{i=1}^{c} f(x|\omega_i,\theta_i)p(\omega_i)} \tag{14}$$

where $P(\omega_i|x,\theta_i)$ is the posterior probability of a new sample x is associated to class ω_i. A new data sample is assigned to the class with the largest posterior probability. With the assumption that the prior probabilities $p(\omega_i)$ of each class are equal (an assumption we make for this two-class example case), the classification rule can be written as

$$\textit{Assign } x \textit{ to class } \begin{cases} 1 & \textit{if } f(x|\omega_1,\theta_1) \geq f(x|\omega_2,\theta_2) \\ 2 & \textit{otherwise} \end{cases} \tag{15}$$

The optimal discriminant boundary for the underlying distributions, given the decision rule of equation (15), is displayed in Figure 8. To compare the discriminant boundary produced by the mixture distribution estimates and the classification rule with the optimal decision boundary, a plot of the decision surface generated by the mixture estimates was computed. The mixture model decision surface is shown in Figure 9.

Given the small number of samples (1000 samples of each class were used for training by Flick et al.), the mixture estimates do well in characterizing the decision surface of the underlying distribution. The optimized generalized mixtures are next compared with the normal component mixture models of their classification capability.

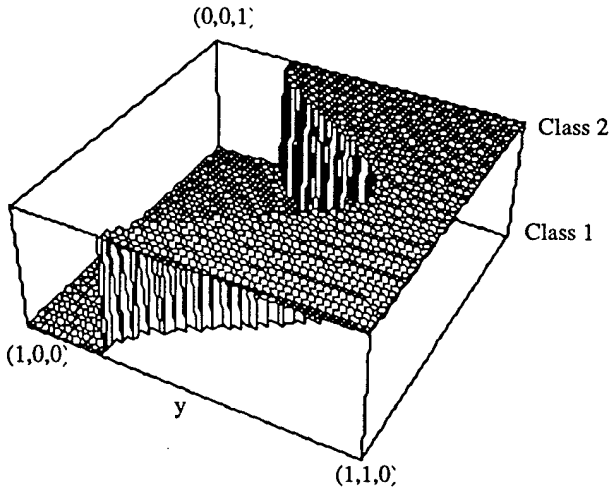


Figure 9. Optimal decision boundary for underlying class distributions.

Classification comparison of the normal component mixture model estimates

To compare the classification characteristics of the two mixture model paradigms, six sets of 5000 samples from each class were randomly generated and classified according to equation (15). The classification performance of the generalized and normal mixture model estimates on these data sets is given in Table 2.

Table 3. Classification performance of generalized and normal mixture models on Flick test data. Numbers correspond to number of correctly classified samples from a test sample of 10000 points (5,000 points from each class).

Data Set	Generalized Mixtures	Normal Mixtures
1	9440	9481
2	9456	9487
3	9425	9466
4	9418	9448
5	9412	9428
6	9452	9464

As noted in the previous section, the theoretically optimal classifier will on average correctly classify 96.50% of the samples presented, due to the overlap of the class distributions. The table demonstrates that both mixture model paradigms produced excellent results, with the generalized component mixture models correctly classifying 94.34% of the test samples versus 94.62% for the normal component mixture models. To test if the classification difference in the mixture approaches is statistically significant, a Wilcoxon paired sample test was applied. For the data in Table 3, a Wilcoxon statistic returns a p-value of 0.05. Therefore a statistically significant difference is detected (at the $\alpha = 0.05$ level of significance) in the classification performance of the two models. Note, however, that this difference is on average only 0.28%, and as demonstrated in the statistics of Table 2, the generalized mixture models are more parsimonious (i.e., they require fewer free parameters to represent the two classes).

6 CONCLUSIONS

Mixture models provide an excellent nonparametric approach for density estimation and pattern classification. The algorithm described by this paper frees a mixture method practitioner from having to make an a priori estimate of the number of components required by a mixture for optimal representation, and allows the practitioner to use shape-parameterized distributions as the functional basis of the mixture components.

This work has demonstrated that the combination of a multiagent stochastic search technique and the EM algorithm can produce sound probability density estimates for multivariate data. The mixture distributions produced by the multiagent optimization process display promising classification capabilities. Although the study is too limited for statements concerning the general classification capabilities of the algorithm and the mixture estimates it produces, this work demonstrates that elegant classifier systems can be produced from algorithmically optimized mixture models.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716-723.

Bäck, T., and H. P. Schwefel (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1:1-23.

Brooks, S. H. (1958), A discussion of random methods for seeking maxima. *Operations Research*, 6:244-251.

Cassella, G., and R. L. Berger (1990). *Statistical Inference*, Pacific Grove: Wadsworth and Brooks/Cole.

Choi, K., and W. G. Bulgren (1968). An estimation procedure for mixtures of distributions. *Journal of the Royal Statistical Society, Series B*, 30:444-460.

Cramer, H. (1946). *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1-39.

Flick, T. E., L. K. Jones, R. G. Priest, and C. Herman (1990). Pattern classification using projection pursuit. *Pattern Recognition*, 23:1367-1376.

Fogel, D. B. (1992). *Evolving Artificial Intelligence*, Ph.D. dissertation, University of California San Diego, La Jolla, CA.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press.

Goldberg, D. E. (1989). *Genetic Algorithms in Search Optimization and Machine Learning*, Reading: Addison-Wesley.

Good, I. J., and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255-277.

Karnopp, D. C. (1963). Random search techniques for optimization problems. *Automatica*, Vol. 1:111-121.

McDonnell, J. R., and D. E. Waagen (1994). Evolving recurrent perceptrons for time-series modeling. *IEEE Trans. on Neural Networks*, 5:24-38.

McLachlan, G. J., and K. E. Basford (1986). *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

Móody J., and C. Darken (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281-294.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065-1076.

Redner, R. A., and H. F. Walker (1984). Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, 26:195-239.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, 14 :1080-1100.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461-464.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, 10:795-810.

Silverman, B. W. (1986). *Density Estimation*, New York: Chapman and Hall.

Solis, F. J., and R. J. B. Wets (1981). Minimization by random search techniques. *Mathematics of Operations Research*, 6:19-30.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, 3:109-118.

Waagen, D., P. Diercks, and J. McDonnell (1992). The stochastic direction set algorithm: a hybrid technique for finding function extrema. In *Proceedings of the First Annual Conference on Evolutionary Programming*, eds. D. B. Fogel and W. Atmar, La Jolla, CA: Evolutionary Programming Society, 35-41.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1995	Professional Paper

4. TITLE AND SUBTITLE

A COMBINED STOCHASTIC AND DETERMINISTIC APPROACH FOR CLASSIFICATION USING GENERALIZED MIXTURE DENSITIES

5. FUNDING NUMBERS

PR: ZW67734A01
PE: 0601152N
WU: DN303002

6. AUTHOR(S)

D. E. Waagen and J. R. McDonnell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Command, Control and Ocean Surveillance Center (NCCOSC)
RDT&E Division
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Chief of Naval Research
Arlington, VA 22217–5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

DTIC SELECTED JUL 20 1995 G

Accesion For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Distribution /
Availability Codes
Dist | Avail and / or Special
A-1

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This work investigates a combined stochastic and deterministic optimization approach for multivariate mixture density estimation. Mixture probability density models are selected and optimized by combining the optimization characteristics of a multiagent stochastic optimization algorithm based on evolutionary programming and the expectation-maximization algorithm. Unlike the traditional finite mixture model, generally composed of a sum of normal component densities, the generalized mixture model is composed of shape-adaptive components. Rissanen's minimum description length criterion provides the selection mechanism for evaluating mixture model fitness. The classification problem is approached by optimizing a mixture density estimate for each class. A comparison of each class's posterior probability (Bayes rule) provides the classification decision procedure. A classification problem is posed, and the classification performance of the derived generalized mixture models is compared with the performance of mixture models generated using normally distributed components. While both approaches produced excellent classification results, the generalized mixture approach produced more parsimonious density models from the training data.

DTIC QUALITY INSPECTED 5

Published in *Evolutionary Programming IV*, March 1995.

14. SUBJECT TERMS		15. NUMBER OF PAGES
Mixture Density Mixture Models	Stochastic Search Neural Network Models	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

21a. NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE *(include Area Code)*	21c. OFFICE SYMBOL
J. R. McDonnell	(619) 553–5762	Code 785